PROJECT DOCUMENTATION

INTERNAL INTERIM PROGRESS REPORT

Project: **BHL-Europe**

Release: Final

Date: 4 August 2009

Author: Henning Scholz

Table of Contents

Internal Interim Progress Report

1   Purpose

To report on a quarterly base the status of work for each member of the consortium. This report will not duplicate the content of the offcial progress and annual reports requested by the EC every 6 and 12 months, respectively.

2   Follow-ups From Previous Reports

This is the first report after the start of the project on 1 May 2009. Therefore, no follow-ups from previous reports are required.

3   Activities

1.) Distribution of library questionnaires to be filled by content providers. In order to collect important data from all content provider of thte BHL-Europe consortium, a questionnaire was distributed end of April 2009. The filled questionnaires were evaluated during the first months of the project and will be the base for the configuration of the pre-ingest procedures.

2.) The grant agreement was received on 30 April 2009 from the EC by the PCO to be distributed to the partners. A-forms were sent out and collected during the Kick-off Meeting and the days after. The signed A-forms were sent back to the EC by 10 June 2009. The grant agreement was signed by the EC and sent to the PCO in early June 2009. After completion of the grant negotiation, the first pre-instalment was received and distributed to the partners in July 2009.

3.) The Kick-off Meeting of BHL-Europe including two initial workshops was hold in Berlin (MfN) from 10 to 15 May 2009. It started with an initial PMG meeting followed by a welcome event in the dinosaur hall of MfN. On 11 May the framework for the project was discussed. 12 May was mainly reserved on external projects that are important for BHL-Europe. Our EC PO was in Berlin on 13 May to introduce some important aspects of project governance, administration and dissemination. On 12 and 13 May all content providers had the opportunity to present their projects in order to start the process of identifying the content holder requirements. The first workshop of the technology experts was hold on 14 and 15 May to identify roles and tasks for WP2 and WP3. The first workshop of the CWG was hold on 14 May to discuss designs and dissemination activities.

4.) Meeting in Paris with Atos/AIT (8 June) and BnF/MNHN (9 June). Several questions concerning tasks and roles in WP3 and the TMB remained unclear after the Kick-off Meeting. Therefore, the meeting with Atos and AIT was necessary to identify the tasks of the key personnel involved. The minutes of that meeting are available on the BHLwiki: https://bhl.wikispaces.com/file/view/minutes_bhl-e_atos_paris_090608.pdf. The key decision of that meeting was to establish a pre-ingest process with tools developed by AIT based on the existing tools of the DISMARC project. Five digital libraries were identified (BHL, BnF, UGOE, LANDOE, NAT) to submit test data (metadata) for the test and adaptation of the tools. Another decisions of that meeting to create an Architectural Design Document (ADD) authored by Lee Namba and follow up with the developments leading to the Duracloud System. Adrian, Roger and

Walter went to Malta to talk to Sandy Payette and others to discuss this topic during the SUN PASIG conference. For the project management of WP3 activities, SCRUM will be used throughout the project.

5.) Data (export files) from BHL (MySql) and OÖ-Landesmuseum (PostgreSql) have been imported into equivalent relational data bases at AIT IT-infrastructure. Data from BNF and Naturalis can be accessed via OAI-PMH access. The mapping of these data into an ESE (Europeana Semantic Elements) compliant schema is under preparation using tools already developed during the DISMARC-project or using ETL-tools (extract-transform-load) available in the open source domain (Pentaho). Results will be presented during the Leiden meeting.

The input bof scanned book accompanied by bibliographic metadata has been tested, using as interim solution the infrastructure of Internet Archive (IA: http://www.archive.org). BHLE-partner CSIC from spain has provided scanned pages, the metadata record has been retrieved the Z39.50 facility of the Exlibris library system available at CSIC premises. The book is alread included in the IA biodiversity collection (http://www.archive.org/details/contritomorpho00marcus). The uploaded and generated files provide input to the BHL-portal which is just underway. Same data from IA will be taken for the BHL-Europe demonstrator in German language. A second book was uploaded to test the processing of Fraktur type face which was popular in the German speaking regions until 1941. These tests are underway (http://www.archive.org/details/bilderatlasmw4). Results of this trial will be presented during the Leiden meeting.

6.) The overall BHL-Europe system architecture has been further elaborated based on the discussions in Berlin, Paris and Malta. As tool for repository management, FedoraCommons was selected. Emphasis is put into three components: pre-ingest (metadata harmonisation and creation, semantic enrichment using ontology WebServices), storage (integrating cloud technology) and access (providing multilinguality and ontology support for query expansion).

The development work of WP3 is supported by the the open source product agile42 which provides web based access to different functionalties used in the SCRUM approach. The necessary IT infrastructure was setup and tested at AIT premises (http://scrum.ait.co.at/projects/bhl); the definition of the initial backlog is underway and at the time of reporting a first sprint has been initiated. As one result of this sprint a draft version of the ADD (architectural design document) will be delivered. The structure of the document follows ESA standards (http://www.esa.int/TEC/Software_engineering_and_standardisation/TECBUCUXBQE_0.html).

7.) The DuraCloud meeting took place at the New York public library on 14th July 2009. Within the workshop details were discussed on (amongst other things) how BHL-A and BHL-E would replicate information between their repositories. BHL-A will move data into a cloud storage environment that incorporates DuraCloud (http://duraspace.org/duracloud.html). Once BHL-A is integrated into the DuraCloud (Uploaded/ingested data) it will be possible to synchronise BHL-A and BHL-E data. The work required to enable the synchronisation by BHL-E is to create a connector that will allow the BHL-E repository to "see" the same cloud space as BHL-A. The Akubra connector (http://www.fedora-commons.org/confluence/display/AKUBRA/Akubra+Project) will provide the bridge between the preservation repository (Fedora Commons in Europe) and the BHL-A storage area (Cloud S3 environment in America). The cost of the Akubra connector development is currently estimated at €10000.

8.) The composition of the Technology Management Board (TMB) was made up from a number of institutions and companies within BHL, the members include:

Kai Stalmann	MfN, Leader WP2
Adrian Smales	NHM, Leader WP3, Chair TMB
Bernard Scaife	NHM
Walter Koch	AIT
Christopher Freeland	MOBOT, BHL
Julie Verleyen	EDL Foundation, Europeana
Anton Guentsch	BGBM, EDIT WP5

The board will have to find a suitable replacement for Bernard Scaife, as he will be leaving the NHM and the BHL project as of 14th August 2009.

The initial meeting of the TMB was conducted on 14th May 2009. The main aims and goals of the project were discussed as well as detailing the deliverables for the next 12 months; owners and contributors were assigned for each of the deliverables within WP2 and WP3. The minutes for the meeting can be reviewed at:

https://bhl.wikispaces.com/file/view/BHL-E_TechWorkshop_all.pdf

9.) The Wroughton Data Centre (AKA Darwin Repository) will provide the long term sustainability for the BHL project and provide a European/Global facility for data access and preservation. The Wroughton Data Centre will provide 4 streams of revenue that can be exploited:

1. Storage for printed/physical media and objects
2. Scanning and OCR
3. Data Centre Services (storage, access, preservation etc)
4. Business Services (Disaster recovery, Business Continuity and Vendor Hospitality)

Each of the 4 areas will be investigated to determine the local (within the UK) market demand and requirement for each. At this stage it is anticipated that the demand will be sufficient to provide the incentive to move forward and build the facility.

Currently, the initial investigations have provoked significant interest from various tier 1 vendors, including Cisco, IBM, EMC, Dell and others. The estimated projection for completion of the facility will be 24 months from starting the build phase. The concept visualisation can be found here: http://www.web3d.co.uk/wroughtondatacentre/

10.) Initiated by Bernard Scaife, discussions were carried out over the BHLwiki to work on the content holder requirements (D2.1). These discussions were based on the BHL specifications. A draft document was released on 24 June 2009 to summarise first round of dicussion. Although a lot of information was collected during the discussions, not all content provider of BHL-Europe took part and provided their view. Only eleven partners actually shared views and information during several weeks of discussions. In a second round of discussion a few more contributions were collected to finish the document on 5 August 2009. Further documents are in preparation as Annexes to D2.1 to add future requirements and to document the implementation of the work carried out based on the requirements.

11.) The preparation of the consortium agreement is going to be difficult. It takes some more time to finish the work and to find the right wording for all parts. However, a preliminary draft will be available for all partners for the BHL-Europe meeting in Leiden on 10 August.

12.) The working team in the National Museum, Prague prepared the deliverables for M3. We worked on corporate design of the project including its logo. The website of the project was designed including a demonstrator. The first number of the newsletter will be ready by the end of M3. In cooperation with Tine Mallerts (Royal Belgian Institute of Natural Sciences) we prepared a

mailing list and a database of relevant conferences. Already in the kick off meeting we delivered the composition of the Communications Working Group. By the end of M3 we will deliver the first dissemination plan.

All deliverables are available as drafts on 31 July 2009. The final drafting and approval will take place in the week 17 August 2009.

13.) We were active in disseminating the project to a wider audience. We were present in two conferences with presentations: "ELPUB 2009 – 13th International Conference on Electronic Publishing" and "Scientific publishing in the Natural History Institutions" (see https://bhl.wikispaces.com/BHL-Europe-slide-library). We have finished some publications and press releases as well (see www.bhl-europe.eu and https://bhl.wikispaces.com/BHL-Europe-publications). We submitted abstracts and papers for forthcoming conferences in various domains (e.g. Annual Meeting of the Paläontologische Gesellschaft, Systematics 2009, LIBER-EBLIDA, Workshop on Advanced Technologies for Digital Libraries 2009).

14.) During the eSciDoc-meeting in Karlsruhe the performance issue of Fedora has been discussed and the use of the eSciDoc system (eventually for publishing) could be one option for the post processing of BHL-Europe data. The representative of Max Planck Society (MPG) was informed about BHL-Europe, since MPG also supports related projects in the field of natural sciences.

15.) EDIT has organised a workshop from 22-23 June 2009: Scientific publishing in the Natural History Institutions. This workshop took place in Bratislava and BHL-Europe was present with a project introduction. The main intention was to discuss aspects of scientific publishing in natural history institutions, opportunities and challenges for electronic publishing and the opportunities for joint publishing. We as a project may take advantage of this forum in the future to get in touch with publishing institutions in order to get permissions to digitise their backfiles for BHL and thus increase our network of content providers.

6 Deliverables and decisions taken

1	D2.1 Catalogue of content holder requirements (quality, quantity, accessibility, standards and specifications of content and metadata) – Final revision 5 August 2009
2	D3.1 Deliver composition of Technology Management Board and initial meeting – 4 August 2009
3	D5.1 Web site, including multimedia presentation – Draft completed 31 July 2009, Final revision in progress
4	D5.2 BHL-Europe newsletter and mailing list – Draft completed 31 July 2009
5	D5.3 Database of relevant conferences/events and ownership for BHL-Europe presentations – Draft completed 31 July 2009, Final revision in progress
6	D5.4 Deliver composition of Communications Working Group and 1st dissemination plan – Draft completed 31 July 2009, Final revision in progress

Decisions taken to date:

- 11 May 2009: constitution of the TMB

- 11 May 2009: dates and locations of next project meetings

- 14 May 2009: consitution of the CWG

- 02 June 2009: design of BHL-Europe logo

- 08 June 2009: develop pre-ingest process using DISMARC mapping tools

- 12 June 2009: layout and structure of project Website

7   Schedule Status/Problems/Issues/Deviations

Most work packages are working almost on schedule. We only have a slight delay with the deliverables that are supposed to be submitted by month 3 of the project (end of July). This delay is partly caused by the delay of the 1st pre-financing. It was particularly difficult for NMP to work on time with no money actually paying people employed for specific tasks. However, no deviations from work plan are required currently or are foreseen for the next phase of the project.
One delay is obvious in WP1: the project management portal is not fully operational yet. Several reasons are responsible for that. First, it was not possible to buy the required software because of the delayed pre-financing. Second, the required data and information from all partners was not in the system in time. However, once the system is operational, a lot of the management work will be much more effective. We expect the system to be running for all partners in the beginning of September 2009.
One important issue for the project in the first months was the lack of the leader of WP2. Although Bernard Scaife was working as the interim WP leader with the support of the PCO to gain momentum, this was not planned this way. However, the recruitment time for the WP leader was to short giving the date of approval of the grant agreement by the EC and the formalities to hire personnel for MfN.
Another issue for the project is the leave of Bernard Scaife by 14 August. This requires a well planned handover of his tasks to the new WP2 leader. This will take place in the first week of August in London.

8   Work Planned

The work planned for the next phase is explained in more detail in Annex 1. WP2 and WP3 have two important milestones in month 6 of the project. By month 6 we have to have consensus about what content from each partner we include in BHL-Europe and how we are going to harvest this content. We also need the first IPR documents in place. This is necessary to start the actual metadata harmonization and content harvest in month 7 of the project in building the BHL-Europe prototype.

9   Products to be Completed in the next phase

7	D1.1 Progress Report 1
8	D3.2 Document agreed standards, best practice and system components
9	D3.3 Plan for managing interoperability issues, data harmonisation and the integration of the content into BHL-Europe, EUROPEANA and the BHL
10	D5.5 BHL-Europe dissemination plan
11	D5.6 BHL-Europe promotion kit